JCST Journal of Data Acquisition and Processing
ISSN: 1004-9037
https://sjcjycl.cn/
DOI: 10.5281/zenodo.7733255

COVID -19 DIAGNOSIS BY USING DECISION TREE ALGORITHM

Zahoor M. Aydam[1], Tasadi M. Hanoon[2] and Ghosoon K. Munahy[3]
[1] Computer sciences and mathematics college, University of Thi_Qar, Iraq
[2]Ministry of education Thi-Qar Education Directorate, Iraq
[3]College of Dentistry, University of Thi Qar, Iraq
Emails: [1] zahoor.mosad@utq.edu.iq , [2] jnattt00@gmail.com and [3] ghosoonalsuraifi@utq.edu.iq

Abstract

The Coronavirus disease 2019 outbreak, began when unidentified pneumonia was discovered in Wuhan, China, and spread worldwide in a matter of days. A long-term negative effect on public health is caused by this outbreak, moreover, So many people lost their lives as a result of it. our study aims to develop an intelligent computer-aided system that can detect positiveCOVID-19 cases automatically, Which may be useful with daily medical problems . In this paper we suggested a system based on image processing and can automatically expose discriminative features on chest X-ray images by followed step that begin with segmentation the lung region then extracted Geometric Moment Invariants (GMI) as feature extractor, decision tree was classified the states. Our result referred to, the Hunt's Algorithm classifier achieved an accuracy of (98.7%), a sensitivity of (0.97, a specificity of (0.98), and an F-score of (0.97). The high detection performance achieved in this study explain the utility of fouse features and a appropriate classifier method for detecting COVID-19 cases in CXR images . With the current resources, this would be greatly helpful in speeding up disease detect.

Keywords: COVID -19, Decision Tree Algorithm, Corona virus

1. **Introduction**

The end of the year 2019 is considered one of the worst years in the current century due to the outbreak of what is known as the Corona virus. It appeared for the first time in the cities of Wuhan, China, and began to spread very widely around the world due to the lack of discovery of any treatment or vaccine that works to reduce the spread (Liu et al., 2017; Zhavoronkov et al., 2020) . Not only did it have effects on the health aspect of the human being, but rather, the movement of the world was stopped from carrying out various activities(Li et al., 2020; Zhang et al., 2020). In 2020 Huang et al. In January, a summary of this virus, as 41 people infected with this virus were studied. Symptoms of this virus are coughing, fever and shortness of breath. These are among the most important symptoms that he is infected with, and some other people do not show symptoms significantly only headache and body weakness. It can spread from one person to another through touch or close contact with an infected person(Huang et al., 2020; Yu, Zhu, Zhang, & Han, 2020). The disease is diagnosed through several methods, one of which is taking a image of the chest area and it is examined by experts, but it takes a long time in case there are hundreds of infected people. By using neural networks to detection the lung region . A database was used to image the lung region. putting Cases into various groups according to their features are called classification. The proceeding underlying this approach a set of rules is formulated according to the training and testing of individual cases. the researchers have been developed several algorithms for classification-based data mining. Among them are decision trees(Gonzalez & Ozguner, 2000).The classification of the results shows that the decision tree algorithm with higher accuracy (99.2%) and sensitivity (96.1%) outperforms the K meen method with an accuracy of 83.4% and sensitivity of 86% (Chen, Zheng, Lloyd, Jordan, & Brewer, 2004). Wang & et al. introduced a system that detects Covid-19 disease by using convolutional neural networks for a set of chest X-ray images. The research worked on image analysis of the affected lungs and normal lungs, where the system's recognition accuracy was 93.3% (Bonchi et al., 2001). Ali & et al. He presented five convolutional neural network models (ResNet50, ResNet101, ResNet152, InceptionV3 and Inception-ResNetV2) on three databases consisting of four categories: healthy lungs, lungs infected with COVID-19 and Viral pneumonia and bacterial pneumonia and the results were as follows: 96.1% accuracy for Dataset-1, 99.5% accuracy for Dataset2 and 99.7% accuracy for Dataset3 (Sharma & Kumar, 2016). Mohammad & et al. They provided a convolutional neural network model to train a set of x-ray images containing three categories normal, pneumonia, and COVID-19. The number of pictures infected with Covid-19 was 180. As the accuracy value for Covid-19 detection reached 99.56%, and the overall rate of all trained items was" percentage(M.-K. Hu, 1962) .

2. **Methodology and method**

2.1. Geometric Moment Invariants (GMI)

Most often, the moment invariant is used to describe the characteristics of objects in digital image processing. There are two kinds of feature extractions based on moment invariants: contour-based or shape-based moment invariants and region-based moment invariants (Park, Lee, & Lee, 1999; Urooj & Singh, 2015). Moment invariants are really advantageous for extracting object features with unique characteristics irrespective of location, size, and orientation. seven invariant moments called as " Geometric Moment Invariants (GMI) " proposed by Hu the Geometric Moment Invariants (GMI) are invariant toward translation, rotation, scaling, and mirroring as follows(Baik & Bala, 2004) .

The geometric moments m*qp*of the order '*z* ' of a 2- dimensional image defined as

$$m_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x^p\, y^q f(x,y)dxdy \qquad [1]$$

Where, ' *p*' and '*q* ' are the non-negative integers such as *z* = *q* + *p* . *f* (*x*, *y*) is 2-dimensional piece-wise continuous real image defined on a compact support *F C Z* * *Z* . The moments defined by eq. (1) may or may not be invariant to rotation, translation, and scaling. The scale invariant is obtained by shifting the image such that it coincides with the origin of the co-ordinates system. The translation invariant is achieved by shifting the polynomial basis into the test image centroid. The resulting moments are called central moments and given as

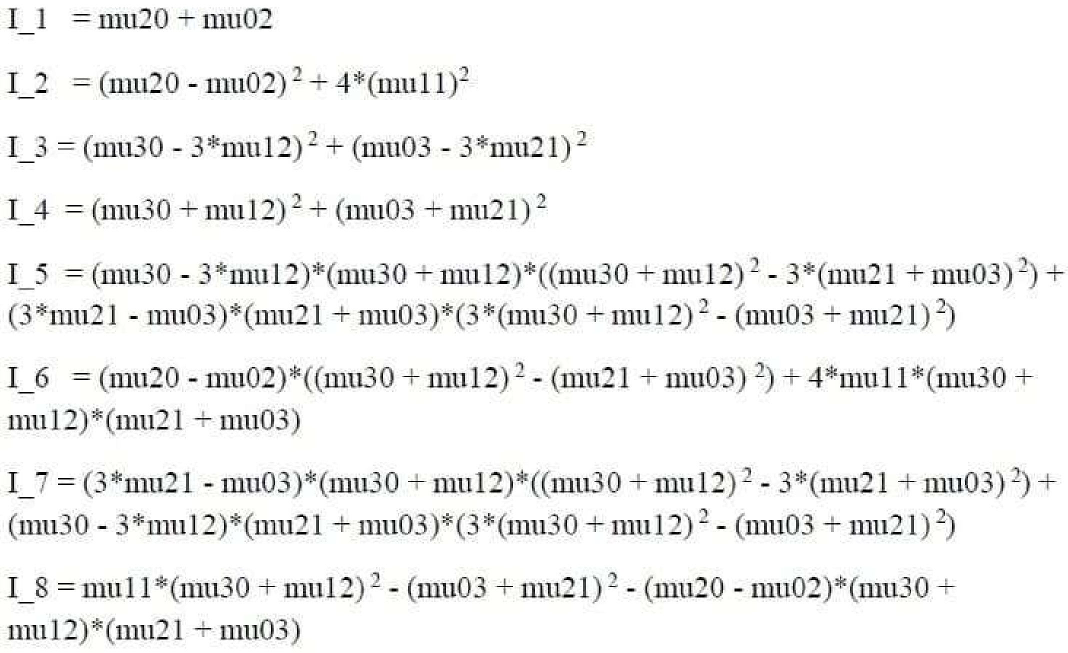
$$\eta_{pq} = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-x')^p\,(y-y')^q f(x,y)dxdy' \qquad [2]$$

Where, *x* ' = *mu*10 / *mu*00 and *x* ' = *mu*01 / *mu* 00. *m*u00 is the mass of the image. *m*u10 / *m*u00 and *mu*01 / *m*00 are the centroids of the image. The scale invariance is obtained by dividing the central moments by a proper normalization factor which is a non-zero quantity for all the test images (MacArthur, Brodley, Kak, & Broderick, 2002).

Any moment can be used for normalization the central moments. It is evident that the lower order moments are more robust to noise and easy to calculate. Therefore, we use the following for the normalization purpose

$$v_{pq} = \frac{\eta_{pq}}{\eta_{00}^{w}} \qquad [3]$$

Where,
$E = ((q + p) / 2) + 1$. In 1962 Hu proposed the following eight invariant moments for in-plane rotation



$I\_1 = mu20 + mu02$

$I\_2 = (mu20 - mu02)^{2} + 4*(mu11)^{2}$

$I\_3 = (mu30 - 3*mu12)^{2} + (mu03 - 3*mu21)^{2}$

$I\_4 = (mu30 + mu12)^{2} + (mu03 + mu21)^{2}$

$I\_5 = (mu30 - 3*mu12)*(mu30 + mu12)*((mu30 + mu12)^{2} - 3*(mu21 + mu03)^{2}) + (3*mu21 - mu03)*(mu21 + mu03)*(3*(mu30 + mu12)^{2} - (mu03 + mu21)^{2})$

$I\_6 = (mu20 - mu02)*((mu30 + mu12)^{2} - (mu21 + mu03)^{2}) + 4*mu11*(mu30 + mu12)*(mu21 + mu03)$

$I\_7 = (3*mu21 - mu03)*(mu30 + mu12)*((mu30 + mu12)^{2} - 3*(mu21 + mu03)^{2}) + (mu30 - 3*mu12)*(mu21 + mu03)*(3*(mu30 + mu12)^{2} - (mu03 + mu21)^{2})$

$I\_8 = mu11*(mu30 + mu12)^{2} - (mu03 + mu21)^{2} - (mu20 - mu02)*(mu30 + mu12)*(mu21 + mu03)$

These moments are invariant not only for rotation but also invariant for translation and scaling.

2.2. Decision tree

several decision tree algorithms have been developed over time. The researchers were attentive to improve its efficiency and ability to deal with various types of data. Below we will discuss some important algorithms.Hunt's Algorithm: the Decision tree in this algorithm is grown by dividing and conquers way. The training records contain more than one class, use an attribute test to partition the data into respectively purer subsets. this algorithm preserves Splitting optimally for each stage based on some threshold value as greedy strategy(Myles, Feudale, Liu, Woody, & Brown, 2004) . Suggested use decision tree classifier to perceptual grouping 3-D features in the aerial image. In A decision tree for content-based image retrieval was proposed. One of the important application for decision tree approaches is the medical research where the Decision tree is most useful in diagnostics of various diseases (Quinlan, 1996). In addition can use it for Heart sound diagnosis.

3. Dataset

A database was used that contains two categories, the first of which is Covid-19 and the other is normal. These images are taken for the chest area. This data was collected from websites on the Internet and part of it from a specialist radiology center. As the number of normal images reached1600, the number of images infected with Covid-19 reached 1000. The base was divided into two parts: training, and testing (S. Hu, Hoffman, & Reinhardt, 2001).

4. Proposed method

As we mentioned in the previous step, the recognition state by using the database that we relied on requires us to divide the data into two groups, training and testing. As show following Fig.1

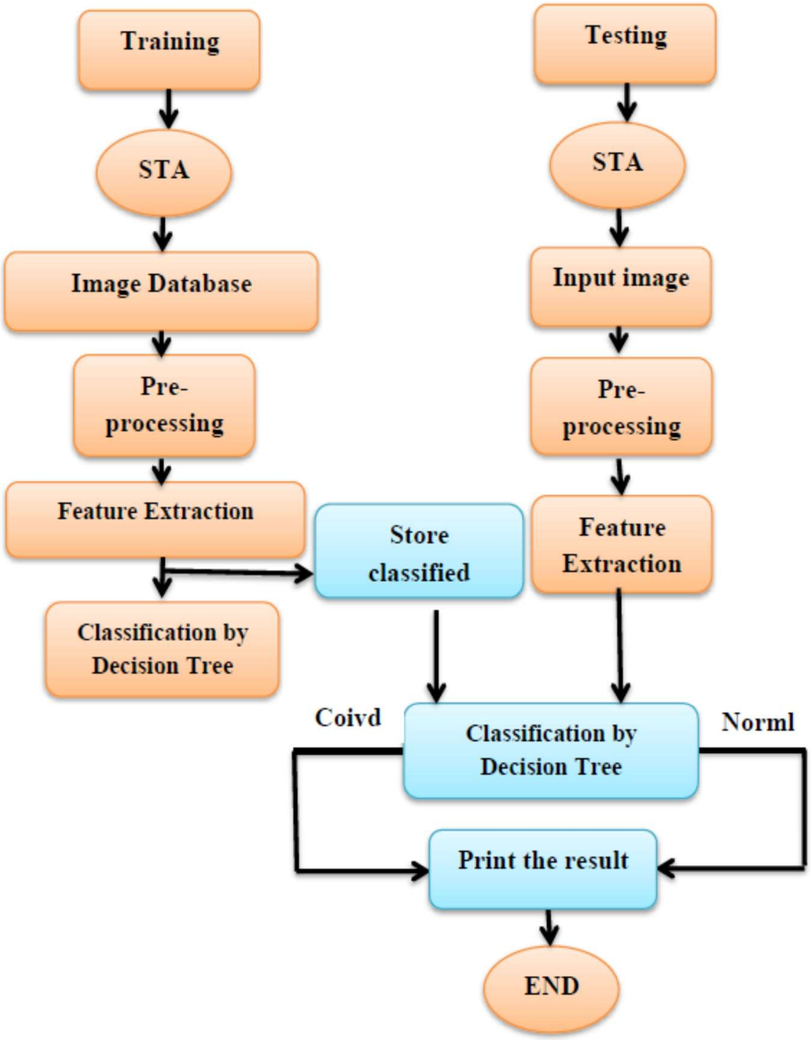


Figure 1. Proposed system

1. Input step: the image is a 3D matrix (width height depth) of values ranging from 0-255. ˆ

2. Pre-processing: At this stage, the images are filtered and noise removed. The images are also made of one size .

a. By using a gray - level thresholding we extracted the lung region from the CT images

b. using a dynamic programming, we parted the left and right lungs by determining the anterior and posterior junctions

c. To achieve results similar to those obtained by manual analysis, we used a succession of morphological functioning to smooth the irregular boundary along the mediastinum, where we excluded the most central pulmonary arteries only from the lung region.

Before pre-processing

After pre-processing

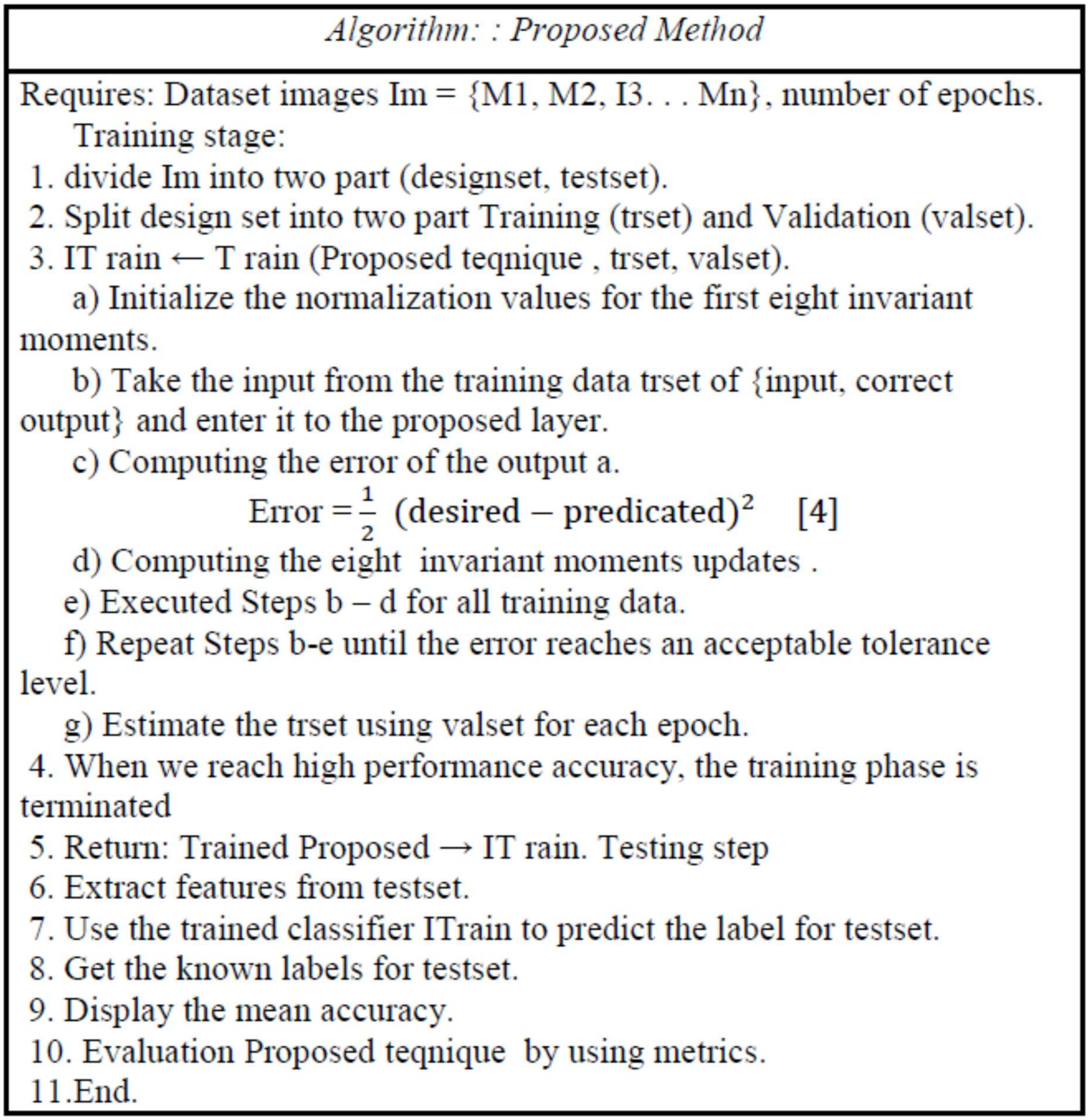
3. Feature extraction step : These step are used for extracted the eight moment invariants . the Moment invariants are really advantageous for extracting object features with unique characteristics irrespective of location.
4. Normalization :the Normalization step reduces the size of activation vectors (vectors for the possibility of using more than one filter). This not only reduces the amount of calculations necessary, but also prevents you from falling into an over fitting state . As show Table. 1.

Table 1. values of invariant moment and normalization it

Type	8 HU moments								Normalization							
Value	XV1	XV2	XV3	XV4	XV5	XV6	XV7	XV8	XV9	XV10	XV11	XV12	XV13	XV14	XV15	XV16
Covid 19	-5.11	-7.84	4.97	-8.29	4.28	2.61	1.35	0.33	-7.78	-1.45	1.06	-5.16	5.27	0.01	0.04	0.47
Normal	-6.09	8.55	-6.42	-8.18	4.24	3.63	4.02	0.57	-8.09	2.85	-3.79	-6.57	5.76	0.02	9.43	0.27

5. Full contacted: After repeating the previous steps several times, all features assemble as a flatten vector until the data enters the final step of the system , which is the fully connected step . The neurons in the two different step s are directly connected to the neurons within the fully connected step as show.
6. Classification step : by decision tree this layer extracts data from the previous step and the difference between the real number and the resulting number from the system is greater, Where the error rate is a measure of the accuracy and efficiency .ˆ
7. Output step: After Classification step and this the last step that carried the normal state or covid state.

5 .Lung recognition: In this stage, By the invariant moments extracts features and saved as vecter. Where images of both normal and infected lungs are trained with Covid-19. Then, they are classified by decision tree according to the characteristics of each base.



Algorithm: : Proposed Method

```
Requires: Dataset images Im = {M1, M2, I3. . . Mn}, number of epochs.
    Training stage:
 1. divide Im into two part (designset, testset).
 2. Split design set into two part Training (trset) and Validation (valset).
 3. IT rain ← T rain (Proposed teqnique , trset, valset).
    a) Initialize the normalization values for the first eight invariant moments.
    b) Take the input from the training data trset of {input, correct output} and enter it to the proposed layer.
    c) Computing the error of the output a.
        $\text{Error} = \frac{1}{2}\ (\text{desired} - \text{predicated})^2$    [4]
    d) Computing the eight invariant moments updates .
    e) Executed Steps b – d for all training data.
    f) Repeat Steps b-e until the error reaches an acceptable tolerance level.
    g) Estimate the trset using valset for each epoch.
 4. When we reach high performance accuracy, the training phase is terminated
 5. Return: Trained Proposed → IT rain. Testing step
 6. Extract features from testset.
 7. Use the trained classifier ITrain to predict the label for testset.
 8. Get the known labels for testset.
 9. Display the mean accuracy.
 10. Evaluation Proposed teqnique by using metrics.
 11.End.
```

6. Performance Evaluation

After building the system and training the database, it must be confirmed that the system is correct and what percentage of it distinguishes infected and non-infected images. In order to achieve the accuracy of the results, he will use a set of measures that determine the number of correctly and incorrectly diagnosed images. Measures such as accuracy, precision, recall, F1-score. A Table 2.

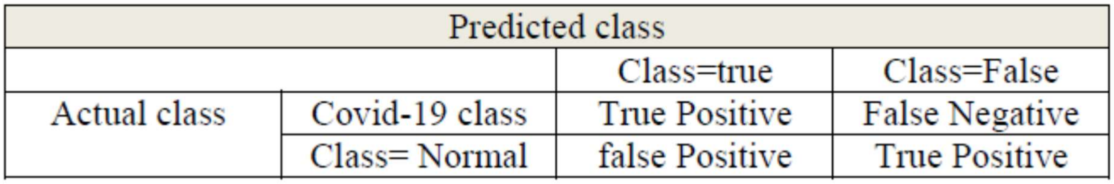
[5] $*100\%$ $\text{Accuracy} = \frac{TB+TN}{TOTAL}$

[6] $*100\%$ $\text{Recall} = \frac{TB+TN}{TP+FN}$

[7] $*100\%$ $\text{precision} = \frac{TB+TN}{TP+F}$

*100% [8] F1 score= $2 * \frac{\text{Recall} \times \text{P recesion}}{\text{P recesion} + \text{Recall}}$

True Positive (TP): correctly that mean the actual class is true and model predicate true.
True Negative (TN): The model predicated negative and classified correctly that mean the actual class is false and model predicate false.
False Positive (FP): The model predicated incorrectly that mean the actual class is false and model predicate true.
False Negative (FN): The model predicated incorrectly that mean the actual class is true and model predicate false.
Table2. Actual Class.



Predicted class			
		Class=true	Class=False
Actual class	Covid-19 class	True Positive	False Negative
	Class= Normal	false Positive	True Positive

7. Performance Experiment

This Experiment was performed by using Covid-19 database, in training state using only 500 normal and 100 covid-19, when the testing state that use 100 normal and 26 covid-19 from each images . The first group contains 600 normal images and the second group contains 126 Covid-19 images .The images were divided into (100 Covid-19 images and 500 normal images) for the training phase and (100 normal images and 26 Covid-19 images) for the testing phase. The accuracy rate in the training phase was 100 and 98.7 in the examination phase as shown in Table 3 below.

Table 3. Accuracy rate

Training phase		Testing phase	
No.image	Accuracy	No.image	Accuracy
600	100%	126	98.7%

8. Conclusion

The results of this research may contribute to improving understanding of dynamics COVID-19 pandemic. We suggest an innovative approach based for a quick and efficient diagnosis of COVID-19 infections, a critical need today. our proposed approach aims to achieve features with contrastive and training the images. Specifically, we use an invariant moment to pick up unbiased feature representations that are strong to overfitting and influence for decision tree for final classification of COVID-19 cases.our proposed method with contrastive loss settings provide a immensely accurate feasible solution for automatically diagnosing COVID- 19 cases to expedite line of treatment for patients. Our best teqnique with deffrents cases achieves an accuracy of 100% in training case COVID-19 cases with impressive values of sensitiv- ity (99.0%) and specificity (98.0%) and 98.7**%** in testing case which are considered to be highly critical performance estimation for applications in medical settings

References

Baik, S., & Bala, J. (2004). *A decision tree algorithm for distributed data mining: Towards network intrusion detection.* Paper presented at the International Conference on Computational Science and Its Applications.

Bonchi, F., Giannotti, F., Manco, G., Renso, C., Nanni, M., Pedreschi, D., & Ruggieri, S. (2001). *Data mining for intelligent web caching.* Paper presented at the Proceedings International Conference on Information Technology: Coding and Computing.

Chen, M., Zheng, A. X., Lloyd, J., Jordan, M .I., & Brewer, E. (2004). *Failure diagnosis using decision trees.* Paper presented at the International Conference on Autonomic Computing, 2004. Proceedings.

Gonzalez, J. P., & Ozguner, U. (2000). *Lane detection using histogram-based segmentation and decision trees.* Paper presented at the ITSC2000. 2000 IEEE Intelligent Transportation Systems. Proceedings (Cat. No. 00TH8493).

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE transactions on information theory, 8*(2), 179-187 .

Hu, S., Hoffman, E. A., & Reinhardt, J. M. (2001). Automatic lung segmentation for accurate quantitation of volumetric X-ray CT images. *IEEE transactions on medical imaging, 20*(6), 490-498 .

Huang, C., Wang, Y., Li, X., Ren, L., Zhao, J., Hu, Y., . . . Gu, X. (2020 .(Clinical features of patients infected with 2019 novel coronavirus in Wuhan, China. *The lancet, 395*(10223), 497-506 .

Li, Q., Guan, X., Wu, P., Wang, X., Zhou, L., Tong, Y., . . . Wong, J. Y. (2020). Early transmission dynamics in Wuhan, China, of novel coronavirus–infected pneumonia. *New England journal of medicine* .

Liu, P., Shi, L., Zhang, W., He, J., Liu, C., Zhao, C., . . . Hu, L. (2017). Prevalence and genetic diversity analysis of human coronaviruses among cross-border children. *Virology journal* , .8-1 ,(1)14

MacArthur, S. D., Brodley, C. E., Kak, A. C., & Broderick, L. S. (2002). Interactive content-based image retrieval using relevance feedback. *Computer Vision and Image Understanding, 88*(2), 55-75 .

Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society, 18*(6), 275-285 .

Park, I. K., Lee, K. M., & Lee, S. U. (1999). *Perceptual grouping of 3D features in aerial image using decision tree classifier.* Paper presented at the Proceedings 1999 International Conference on Image Processing (Cat. 99CH36348).

Quinlan, J. R. (1996). Learning decision tree classifiers. *ACM Computing Surveys (CSUR), 28*(1), 71-72 .

Sharma, H., & Kumar, S. (2016). A survey on decision tree algorithms of classification in data mining. *International Journal of Science and Research (IJSR), 5*(4), 2094-2097 .

Urooj, S., & Singh, S. P. (2015). *Rotation invariant detection of benign and malignant masses using PHT.* Paper presented at the 2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom).

Yu, P., Zhu, J., Zhang, Z., & Han, Y. (2020). A familial cluster of infection associated with the 2019 novel coronavirus indicating possible person-to-person transmission during the incubation period. *The Journal of infectious diseases, 221*(11), 1757-1761 .

Zhang, S., Wang, Z., Chang, R., Wang, H., Xu, C., Yu, X., . . . Cai, Y. (2020). COVID-19 containment: China provides important lessons for global response. *Frontiers of Medicine*, 1 .

Zhavoronkov, A., Aladinskiy, V., Zhebrak, A., Zagribelnyy, B., Terentiev, V., Bezrukov, D., . . . Orekhov, P. (2020). Potential COVID-2019 3C-like protease inhibitors designed using generative deep learning approaches. ChemRxiv. *Preprint posted online on February, 11* .